Randomization

Tim Hanson

Department of Statistics
University of South Carolina

January, 2017

Modified from originals by Gary W. Oehlert

2 Randomization vs. haphazard

An experiment is **randomized** if the method for assigning treatments to to units to units involves a known, well-understood probabilistic scheme, called a randomization. Aspects besides treatments may also be randomized.

Haphazard is not random.

Mice in a cage

Consider assigning treatments A (water) or B (water + kava extract) to 20 mice that just arrived from the Jackson Laboratory; 10 should go into each treatment group. Here are some methods for assigning treatments:

- Give the 10 rats closest to the box opening A, and the other 10 B.
- Give the slowest rats A (i.e. those caught first), and the faster ones B.
- Place 10 red balls and 10 green balls into an urn. Take each rat out in turn and draw a ball; red = A and green = B.

Pros/cons of each?

Why randomize?

Randomization protects agains confounding.

Randomization can be a basis for inference.

Bypass surgery example (p. 15)

Bypass surgery is major; patients w/ severe disease may not survive the operation. A doctor may be tempted to assign stronger patients to surgery & weaker patients to drug therapy.

This confounds strength of the patient with treatment differences. The drug therapy would likely have a lower survival rate because it is getting the weakest patients, even if the drug therapy is every bit as good as the surgery.

Randomization obviates confounding

On average, randomization balances assignments of treatments to subjects.

- Each treatment gets approximately the same number of men vs. women.
- Each treatment gets approximately the same number of older subjects.
- Each treatment gets approximately the same number of weaker subjects.
- Each treatment gets approximately the same number of professional birthday party clowns and/or mimes.

And there's more!

We don't even need to know what we should balance for; randomization does it (approximately) for us.

The deviation implied in the "approximately the same" assignments follows an understood probability mechanism that we can account for.

Without randomization, deviation from balancing can lead to confounding: the units are different in some unknown way, and we cannot account for it.

Lack of randomization can cause big trouble.

2.3 Example in R

To randomly assign N units into groups of size $n_1, n_2, \ldots, n_g$ with $n_1 + n_2 + \cdots + n_g = N$, first put the units in random order. Take first n_1 of the randomly ordered units for group 1, and so on. Alternatively, sample unit/subject labels from a bag without replacement. Here's both in R:

```
perm=sample.int(20)
treatA=perm[1:5]
treatB=perm[6:10]
treatC=perm[11:15]
treatD=perm[16:20]

labels=1:20
treatA=sample(labels,5) # take 5 labels without replacement
treatB=sample(labels[-treatA],5) # take 5 from remaining 15
treatC=sample(labels[-c(treatA,treatB)],5) # 5 more
treatD=labels[-c(treatA,treatB,treatC)] # what's leftover
```

There are other functions/packages in R to do weighted sampling, etc.

One treatment having two levels

We want to see how a response changes with two levels, e.g. relief time from placebo vs. a new allergy drug; heart rate before vs. after caffeine. Two ways to do this:

- Collect N individuals, randomly allocate them into two groups (say placebo or allergy drug) of sizes n_1 and n_2 where $n_1 + n_2 = N$.
- Collect N individuals, each of whom will get both placebo and the drug; randomize the order in which the individuals get either placebo or drug. Need to wait a certain amount of time before applying the 2nd level of treatment to minimize **carryover effects**.

First experiment results in a **two-sample analysis**, second in a **paired** analysis. Paired analysis example of **repeated measures** and **blocking** (more later).

Note for the caffeine example, treatment orders are not randomized! Interest is in how caffeine affects resting heart rate after ingestion; so heart rate measured before and after caffeine. Analysis is still paired though.

Two-sampled tests and paired tests

Read my STAT 205 notes on two-sample hypothesis tests here:
http://people.stat.sc.edu/hansont/stat205/lecture12.pdf
http://people.stat.sc.edu/hansont/stat205/lecture13.pdf

Paired tests are covered here:
http://people.stat.sc.edu/hansont/stat205/lecture16.pdf

Reviews ideas in Prof. Oehlert's Chapter 2 in more detail; self-contained.

For two sample inference, my notes use $Y_{11}, \ldots, Y_{1n_1}$ as responses from level one and $Y_{21}, \ldots, Y_{2n_1}$ level 2. Sample means are $\bar{Y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}$ & $\bar{Y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{2j}$ and variances are $s_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2$ & $s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2$. Finally, $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Note $E(\bar{Y}_1) = \mu_1$ and $E(\bar{Y}_2) = \mu_2$.

2.4 Randomization for inference

Usual parametric inference w/ normal distributions, e.g. t-test:

- Data are assumed to be random samples from some distribution (often normal).
- Inference is about parameters (usually means) of the distributions, e.g. μ_1 and μ_2.
- Sampling induces a distribution on the test statistic (under the null hypothesis), e.g. the t distribution with some df.
- P-value says something about how far the observed test statistic is into the tails of the distribution. Typically reject if p-value < 0.05.

Groups are known, data are random samples.

Randomization testing turns that on its head:

- The null is that the treatment groupings are meaningless and do not affect the responses. The data are considered fixed.
- The only thing random is the assignment of the data to groups.
- Randomization induces a null distribution on a test statistic.
- P-value says something about how far the observed test statistic is into the tails of the distribution.

Data are known (constant); groups are random.

Permutation testing

1. Two-sample version of randomized test called a **permutation test**.
2. Computing the p-value exactly requires enumerating all possible ways to allocate data into groups of size m and n, i.e. all permutations.
3. The results of randomization tests and standard tests are often quite close, but not always.
4. Most of our analyses will assume normal outcomes but this not assumption not always reasonable...

Where is the value in randomization tests?

> If you did the randomization, then the randomization test is valid.

No debates about assumptions, residuals, and so on. It just works. That can be important in legal settings. Also provides more power to reject H_0 if data are very unlike normal distributions.

Two-sample setting

Data $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$. We are testing for equal means. The (pooled) two-sample t-test is

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{1/m + 1/n}}$$

with

$$s_p^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{j=1}^{m} (y_j - \bar{y})^2}{n + m - 2}$$

Null distribution is t with $n + m - 2$ df. This assumes normality, independence, and equal variances. (There is also an unpooled version of the test, but the same points apply.)

With the same data, the randomization test statistic is

$$t = \bar{x} - \bar{y}$$

The null distribution is found by computing the difference of means for all possible assignments of N units into groups of size m and n (i.e., all possible randomizations). The only assumption is the randomization.

These outcomes are equally likely. The p-value is fraction as extreme or more extreme than statistic in the data.

${}_N C_n = \frac{N!}{n!(N-n)!}$ grows very quickly with N, making an exact computation cumbersome.
${}_{10}C_5 = 252$ ${}_{20}C_{10} = 184,756$ ${}_{30}C_{15} = 15,511,752$. Don't want to do it by hand!

Paired setting

In paired data we have x_i and $y_i, i = 1, 2, \ldots, n$ measured on the same unit or units that are similar in some way. Inference is on the differences $d_i = x_i - y_i$.

The paired t-test is

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

with

$$s^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2/(n-1)$$

Null distribution is t with $n - 1$ df. This assumes normality, independence, and constant variance.

With the same data, the randomization test statistic is just $\bar{d}$.

Under the randomization, the two units in the pair either received treatments A and B (in that order), or B and A (in that order). Under the randomization null hypothesis, the sign of each difference plus or minus with probability one half (independently). Called the **sign test**.

The null distribution is found by looking at $\bar{d}$ for all 2^n possible outcomes for the n different signs.

These outcomes are equally likely. The p-value is fraction as extreme or more extreme than statistic in the data.

Paired experiment example w/ R code

Bezjak and Knez (1995) give data on seconds it takes $N = 30$ garment workers to runstitch a collar on a mans shirt, using a standard workplace and a more ergonomic workplace. Are times the same on average for the two workplaces? Let $E(d_i) = \mu$, the mean difference in standard vs. ergonomic. Want to test $H_0 : \mu = 0$ vs. $H_0 : \mu > 0$.

```
library(oehlert) # contains all data sets from G. Oehlert’s book
library(cfcdae) # contains function permsign.test()
emp02.1 # look at data; Table 2.1 in book (p. 20)
d=emp02.1$std-emp02.1$ergon # compute differences
d # look at differences
summary(d) # summary statistics
hist(d,frequency=F) # data approximately normal?
shapiro.test(d) # formal test
t.test(d,alt="great") # one-sided t-test
permsign.test(d,plot=TRUE) # permutation test
```

Two-sample experiment example w/ R code

Hunt (1973) provides experimental data on absorption of phosphorus by Rumex acetosa at 15 and 28 days after first harvest. These are $N = 8$ plants randomly divided into two groups of $n_1 = n_2 = 4$. Is average log-phosphorus content the same or does additional growing time accumulate more phosphorus? Let $E(x_i) = \mu_1$ (15 days) and $E(y_i) = \mu_2$ (28 days); want to test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 < \mu_2$.

```
library(perm) # has two-sample permuation test function permTS()
x15=c(4.3,4.6,4.8,5.4); y28=c(5.3,5.7,6.0,6.3) # Table 2.4 (p. 25)
# can also type library(oehlert) and use emp02.2
log.phos=c(x15,y28); group=c(15,15,15,15,28,28,28,28)
boxplot(log.phos~group) # side-by-side boxplots
t.test(x15,y28,alt="less") # UNPOOLED two-sample t-test
permTS(x15,y28,alt="less") # two-sample permutation test
wilcox.test(x15,y28,alt="less") # Wilcoxin-Mann-Whitney test
```

The Wilcoxin-Mann-Whitney test does not use the sample means as test statistics and is an alternative nonparametric test that does not assume anything about the two distributions. See `http://people.stat.sc.edu/hansont/stat205/lecture15.pdf` for more details if interested.

If you cannot get the cfcdae package to load

The code that defines this function can be run in R using the following:

```
source("http://people.stat.sc.edu/hansont/stat506/permsign.test.R")
permsign.test(d,plot=TRUE) # permutation test
```

Sign test on the median

Instead of looking at the mean difference, we can also look at the median difference; this is what is in Lecture 16 in my STAT 205 notes. The median is not affected by outliers to the same degree the mean is, so if data are highly skewed the median may be more natural to look at. For example median income is more indicative of what a typical person makes than mean.

We have $d_i = x_i - y_i$. Let η be the median difference between treatments A and B in the paired setting. Testing $H_0 : \eta = 0$ is especially easy, as shown below.

```
binom.test(sum(d>0),length(d)) # two-sided alternative
binom.test(sum(d>0),length(d),alt="great") # alternative H0: mu1>mu2
```

Note that the p-values are smaller than when testing the mean! In fact, the one-sided alternative has a p-value below 0.05 using the median. Food for thought...